Jakatón 2015: Report

April 29th and 30th

The goal of the hackaton is to motivate students to get a hands on experience on natural language processing techniques and maybe to attract them to masters of PhD studies in our field. In this respect we were able to bring together to nine students from different universities and institutions from Mexico. Five projects were developed in different aspects of NLP on twitter data, these projects were supervised and evaluated by members of the NLP community in Mexico from three mexican research group (Grupo Golem-UNAM, LTN-INAOE, LKE-BUAP) and one from France (RCLN).

Participants

In this edition twelve candidates registered from which nine participated in the event. They came from six institutions, five which were universities, and one which was a not gubernamental organization focus on open data for governments. The participants come from five different states of the country and they were 4 undergraduates, 2 masters, 1 PhD students; two professionals one with a PhD and other with a career in data mining. From this group seven persons were associated to an NLP laboratory in which they do or have an antecedent with natural language processing and two which were facing NLP problems in their work. The following table presents the list of the participants.

Fernando Monroy Tenorio	IPN, Distrito Federal
Miguel Salazar	ITESO, Guadalajara
Ricardo Daniel Alanís Tamez	Codeando, Nuevo León
Jorge Martínez Ortega	Codeando, Guadalajara
Albert Manuel Orozco Camacho	UNAM, Distrito Federal
Angel Callejas	UAM, Estado de México
Aarón Ramírez De la Cruz	UAM, Estado de México
Fernando Aguilar Reyes	ITAM, Distrito Federal
José Eduardo Casillas García	UNAM, DIstrito Federal

Projects

The participants from five teams which proposed the following projects:
- ***Detection of violence towards woman on twitter*** :This team proposed a NLP pipeline to detect tweets which expressed violence. Among the techniques exploited were: word list, POS tagging and pattern matching. (Team: Fernando Aguilar, Miguel Salazar and Jorge Martinez)
- ***Detection of expressions related to drug consumption:*** This team proposed a NLP pipeline to extend a seed dictionary with expression which refer to the consumption of drugs. Among the techniques they focused on word list and mutual information metrics (Ricardo Alanis).
- ***Sentiment analysis for twitter***: This team used sentiment analysis to analyse the sentiment of the states of México. They based their approach on word lists (Fernando Monroy).
- ***Clustering of tweets*** This team tried to cluster tweets given their distance. To deal with the large vocabulary they choose to replace synonyms word by a word and user a weighted Sørensen distance, once clustering they used a hand written modus ponens rules to consolidate clusters (Jose Casillas y Albert Orozco).
- **Sentiment analysis of media twitter accounts** Finally this team used sentiment analysis to analyse some media twitter accounts and their vocabulary. They used a list of words for categorizing tweets on five basic emotions. They analysed a set of tweets from different media outlets tweets (Ángel Casillas y Aaron Ramirez).

Given the available time for developing the project all were at a stage of proof of concept. In particular, two of the NLP pipelines were not finished. However, during the assessment the project members of the teams discussed what would they need to complete their projects. The other three projects presented a finished pipeline, however only one team was able to provide full evidence of the reach of the obtained results, this was the team of sentiment analysis for the media twitter accounts, this together with their methodology, use of the tools and the examples presented made the judges declare them project the team with the best project of the hackatón.

Program

The hackaton was split on two days: Wednesday, April 29 and Thursday, April 30. Paricipants were programming 8-10 hours per day on the robot hall from the Golem group, at IIMAS - UNAM. There was a social program for the jury (a visit to the Frida Kahlo museum) and a gala dinner for judges and participants. The winner team (sentiment analysis of media tweets) received two iPads as a prize. There was another finalist team (clustering of tweets). The decision was split between the three judges because of the good quality of both finalists. Both finalist teams were composed of college students (the finalists from UNAM, the winners from UAM university).

Resources

Some of the projects can be found in our main repository:
- Violence towards woman: https://github.com/jakaton/jakaton_feminicidios
- Clustering of tweets: https://github.com/jakaton/twittemantics
- Sentiment analisis media: https://github.com/jakaton/PerfilNoticieros

Besides the organizers made available the following tools:
- Web service for POS tagging based on Core NLP Standford:
  https://github.com/jakaton/pos_flask
- Web service for sentimenta scoring based on list:
  https://github.com/jakaton/listscore_flask
- Basic library for NLP https://github.com/jakaton/jakatonlib

Website:
- http://golem.iimas.unam.mx/jakaton

Funding

Funding for the event was obtained through NAACL Emerging Regions Funding (19,500 MNX, 1,500 USD) and INAOE projects (30,000 MNX). The resources were used for the following expenses:

Travelling	18,400 MNX
Accomodation	6,996 MNX
Food	8,322 MNX
Prices	8,000.00 MNX
T-shirts	2,400.00 MNX
Administrative fee	3,500 MNX
Total	47,620 MNX

Organizers

- Dr. Ivan Vladimir Meza Ruiz

  Grupo Golem, IIMAS-UNAM

- Dr. Jorge Garcia Flores

  LIPN - Université Paris 13 / Grupo Golem, IIMAS-UNAM

- Dr. Luis Villaseñor

  Laboratorio de Tecnologías del Lenguaje, INAOE
- Dr. David Pinto

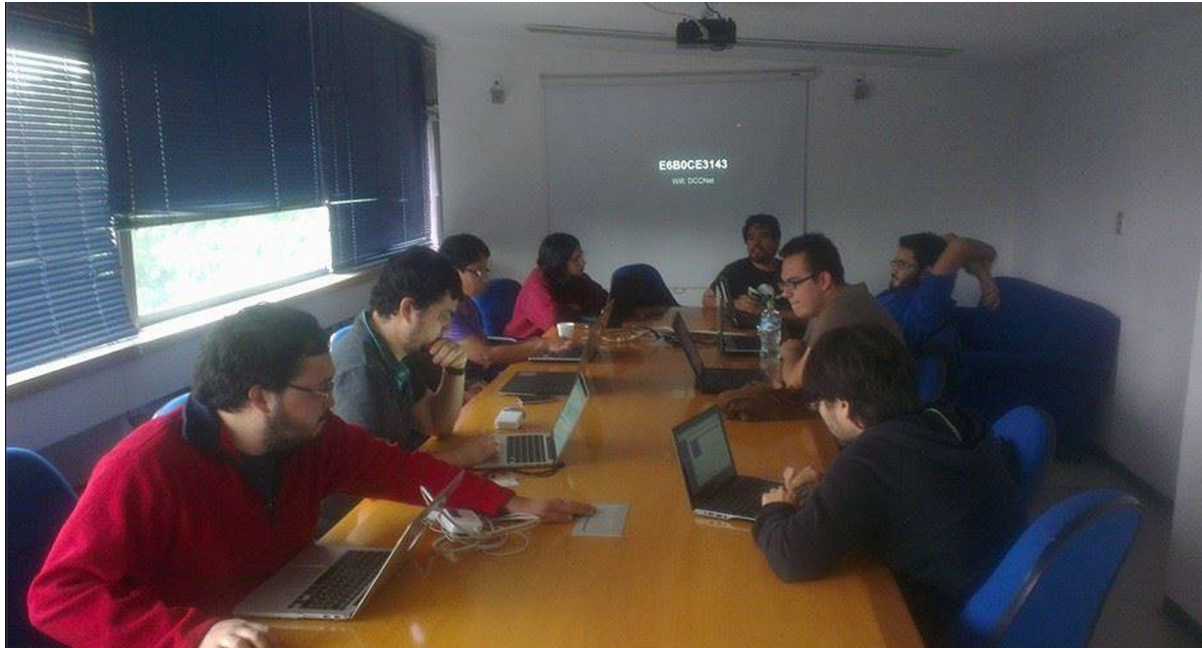 Ingeniería del Lenguaje y del Conocimiento

Institutions

- Grupo Golem, DCC-IIMAS-UNAM(CEDE)
- Laboratorio de Tecnologías del Lenguaje, INAOE
- Ingeniería del Lenguaje y del Conocimiento,BUAP
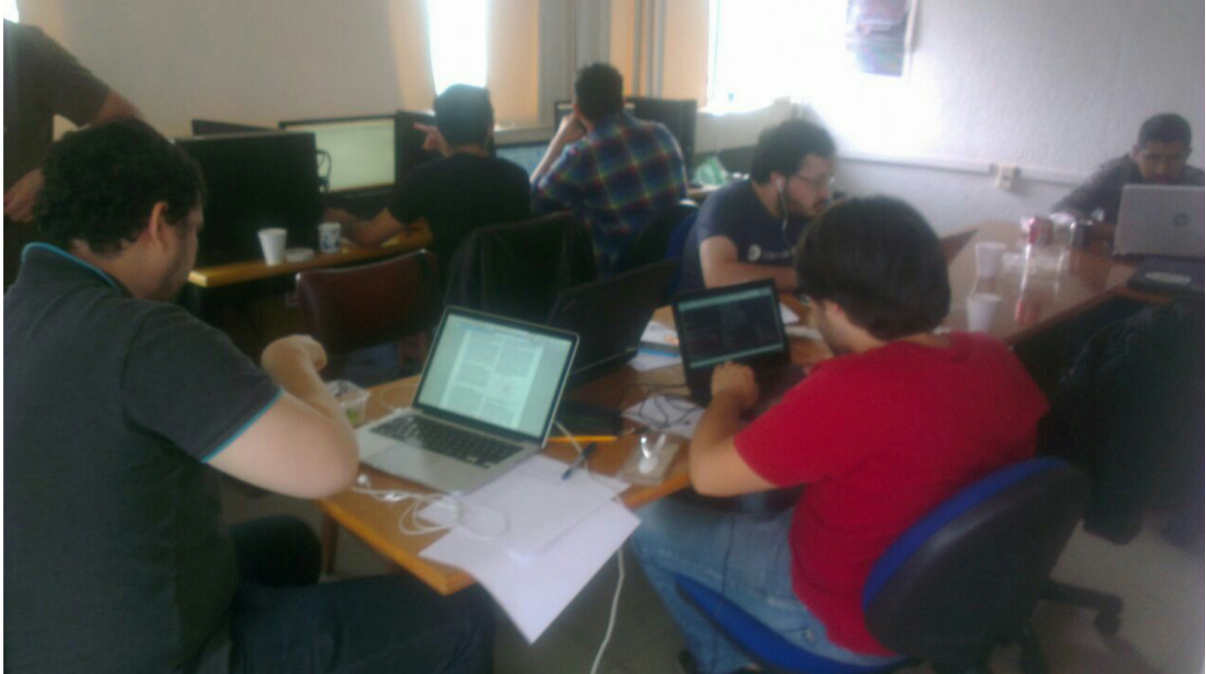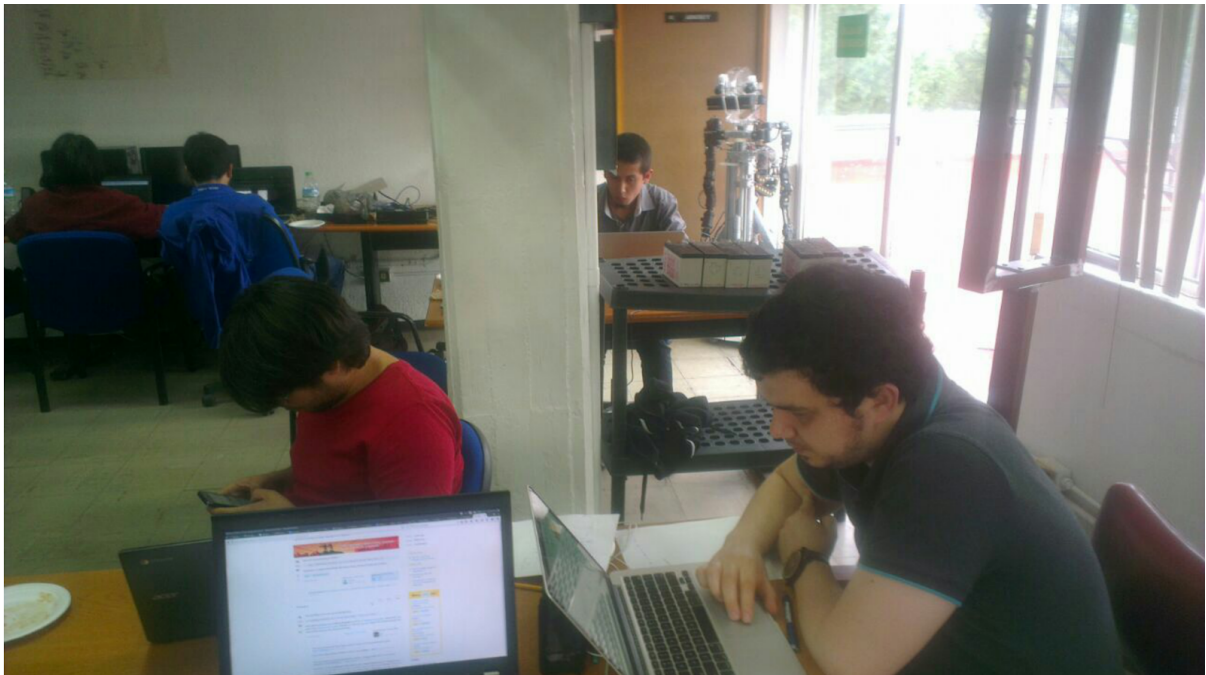- LIPN - Université Paris 13 / CNRS

Acknowledgments

Pictures

Hackaton kickoff

Concentration

Programming in the robot hall

Finalists

Winners

Sunset with winners and finalists

Gala dinner